Journal of Theoretical and Applied Information Technology
15th December 2025. Vol.103. No.23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

FACEDEPAI: A DEEP LEARNING FRAMEWORK FOR DEPRESSION DETECTION FROM FACIAL EXPRESSIONS USING FACEDEPNET

MD ZAINUDDIN NAVEED[1*] AND DR. SHIVAMPETA APARNA[2]

[1]Research Scholar, Department of Computer Science and Engineering GITAM-School of Technology, Hyderabad. GITAM (Deemed to be University), Hyderabad Campus Rudraram, Patancheru Mandal, Sangareddy Dist., Hyderabad, Telangana 502329.

mnaveed2@gitam.in

[2]Associate Professor, Department of Computer Science and Engineering GITAM-School of Technology, Hyderabad. GITAM (Deemed to be University), Hyderabad Campus Rudraram, Patancheru Mandal, Sangareddy Dist., Hyderabad, Telangana 502329.

ashivamp@gitam.edu

ABSTRACT

Depression is a widespread mental health problem with a significant effect on productivity and quality of life. Early, objective detection is still a crucial problem as existing diagnostic methods like clinical interviews and self-reported questionnaires are subjective and resource demanding. Advances in artificial intelligence have facilitated the construction of an automatic depression detection system based on multiple modal data signals, such as text, speech, and image. However, the current image-based methods may overfit to small datasets and not focus on depression-specific characteristics with poor interpretability, leading to their limited effectiveness in clinical applications. To address the gaps above, we propose presenting FaceDepAI, a deep learning system designed for binary depression detection from facial expressions, based on our model, FaceDepNet. The architecture combines convolutional neural networks with squeeze-and-excitation attention modules to dynamically reweight feature channels, emphasizing subtle yet clinically significant facial cues. We use data augmentation and dropout regularization to improve generalization, while incorporating Grad-CAM for increasing interpretability by highlighting depression-related regions on the face. The framework was tested on the DAIC-WOZ dataset with five-fold cross-validation. The experiment results showed that our proposed method achieved 96.8% accuracy, 96.5% precision, and a recall of 97%, with an F1-score as high as 98.2 % ROC-AUC, which consistently outperformed the baseline models, including ResNet-18, VGG-Face, and EfficientNet without an attention mechanism. These results demonstrate FaceDepAI is a robust, explainable, and clinically meaningful tool for depression detection. The scale can be used for broader, larger-scale mental health screening and as an aid to clinical decision making.

Keywords - *Depression Detection, Facial Expressions, Deep Learning, Attention Mechanism, Explainable AI*

1. INTRODUCTION

Depression is one of the most prevalent mental health disorders worldwide, affecting more than 280 million people and contributing significantly to global disability and reduced quality of life. Early and reliable detection of depression is critical, as timely diagnosis can improve treatment outcomes and reduce the risks of chronic mental illness or suicide. Traditional diagnosis primarily relies on clinical interviews and self-reported questionnaires such as the Patient Health Questionnaire (PHQ-8), which, while widely adopted, are subjective and limited by reporting bias 22. This has led to increasing interest in artificial intelligence (AI) and deep learning methods for automated depression detection, which provide scalable and objective screening solutions.

Prior studies have explored multimodal approaches, including textual data from social media posts 33, acoustic features from speech recordings 44, and visual features from facial expressions 55. Text-based models often capture semantic indicators of depressive language, while

speech-based models leverage prosodic variations such as tone and pitch. Visual modalities, though less extensively studied, are particularly valuable, as facial cues such as reduced expressivity, gaze aversion, and downward lip curvature often provide non-invasive biomarkers for depressive states. Existing visual approaches, however, face limitations such as overfitting on small datasets, inadequate focus on depression-relevant features, and limited interpretability of model outputs 66. These challenges highlight the need for novel deep learning frameworks that combine feature refinement with explainability for reliable clinical integration.

The objective of this research is to develop an end-to-end framework, termed FaceDepAI, capable of detecting and classifying depression from facial images using a novel deep learning model, FaceDepNet. The key novelties of this work include the incorporation of squeeze-and-excitation attention modules to reweight feature channels dynamically, thereby emphasizing clinically relevant facial patterns; the use of data augmentation and dropout regularization to enhance generalization; and the integration of Grad-CAM for visual interpretability of predictions. These design choices collectively address significant limitations of prior works and establish FaceDepAI as a robust, explainable, and clinically meaningful system.

Despite significant advances in deep learning for emotion and mental health analysis, current depression detection systems continue to face major challenges related to generalization, interpretability, and clinical scalability. Existing image-based frameworks frequently overfit to small, homogeneous datasets and provide limited transparency in decision-making, restricting their acceptance in real-world clinical contexts. Therefore, this research addresses the following problem statement: to design and validate an explainable deep learning framework that accurately detects depression from facial expressions while ensuring interpretability and generalizability across subjects. Guided by this goal, the study pursues four specific objectives: (1) to develop a customized CNN architecture, FaceDepNet, enhanced with squeeze-and-excitation (SE) attention for refining depression-specific features; (2) to implement data augmentation and dropout regularization for robust learning; (3) to incorporate Grad-CAM-based explainable AI for visual interpretability of decisions; and (4) to evaluate performance through rigorous five-fold cross-validation on the DAIC-WOZ dataset. These objectives collectively align the proposed work with the identified gaps in current literature and establish the foundation for the FaceDepAI framework.

The contributions of this research are multifold. First, a tailored CNN-based architecture with attention integration is proposed for depression detection from static facial frames. Second, extensive experimentation on the DAIC-WOZ dataset demonstrates significant improvements over baseline models such as ResNet-18, VGG-Face, and EfficientNet without attention. Third, ablation studies quantify the contribution of attention, augmentation, and dropout, confirming their collective impact on performance. Fourth, explainability through Grad-CAM bridges the gap between model predictions and clinical interpretability, strengthening trust and adoption potential.

The remainder of the paper is structured as follows: Section 2 goes through related work in AI-based depression detection among modalities. The proposed approach is described in Section 3, which consists of dataset generation, the FaceDepAI framework, and its corresponding architecture (FaceDepNet). Experimental results (setup, performance evaluation, and baseline comparison) along with an ablation study and explainability validation are described in Section 4. Section 5 interprets the findings and presents limitations to this study. Section 6 concludes the paper and suggests directions for future work.

2. RELATED WORK

Depression detection has received considerable attention as a research problem in recent years, with a strong emphasis on developing techniques based on artificial intelligence (AI) and computer vision for analyzing facial expressions as objective markers of mental health states. Several studies indicate that facial dynamics carry subtle signatures of depressive symptomatology, overlooked by the majority of conventional methods for diagnosis. For instance, Rajawat et al. [1] proposed a fusion of fuzzy logic and deep learning to enhance depression diagnosis performance, Cao et al. [2] presented a systematic literature review of deep learning approaches for facial expression analysis and described the challenges such as small datasets, overfitting, and lack of clinical interpretability. Other researchers

have also focused on analyzing micro-expressions and movement dynamics through deep learning to describe psychological changes of depression [3].

Early work on facial emotion and depression recognition developed a pipeline using CNNs; however, they are limited by small dataset size and poor generalization. Neha et al. [4] investigated CNNs for simultaneous emotion recognition and depression detection, and Lee et al. [5] presented the emotional profile of depressed patients using facial dynamics analysis in the time domain. Similarly, Lee and Park [6] introduced a fast R-CNN method for depression diagnostics based on facial expressions, which was demonstrated to be beneficial in terms of region-based feature extraction. More general applications of deep learning techniques in stress and emotion recognition were also confirmed by Almeida and Rodrigues [7] as well as Pise et al. [8]. Agung et al. [9] further demonstrated the effectiveness of CNNs in the field of image-based recognition, showing state-of-the-art performance on large-scale datasets. In the field of mental health, Liu [10] and Tufail et al. [11] demonstrated the effectiveness of CNNs for mental health assessment associated with depression. They proved that their approach applied to various populations.

Recent innovations introduced transfer learning and attention mechanisms to improve detection performance. Publications such as Alsadhan [12] and Hafiz et al. [13] demonstrate the improvement in feature representation achieved with pre-trained models. Bhagat et al. [14] also successfully utilized CNNs for FER, yielding good results. Hybrid methods were also proposed, i.e., the use of concatenated image representations in sleep apnea detection [15] or transformer augmented CNNs for malware classification [16]. Methodological inspiration can be drawn from such efforts in depression detection. Krause et al. [17] reported a meta-analytic review of facial affect recognition in depression, supporting their claim that decreased positive and increased negative facial expressions characterize depressed people. Complementarily, Fu et al. [18] utilized facial emotions mimicry to diagnose depression, and discovered the behavioral markers that are of diagnostic significance; others, Gao et al. [19], demonstrated overlapping emotional recognition challenges between depression and schizophrenia.

From a clinical perspective, Li et al. [20] proposed a behavior-based depression diagnosis model incorporating emotional conflict, bridging AI with cognitive psychology. Broader surveys, such as Khan [21] and Chowanda [22], analyzed the strengths and limitations of deep learning approaches for emotion recognition, underscoring challenges of data imbalance and model interpretability. Expanding beyond facial data, Aleem et al. [23] and Squires et al. [24] surveyed machine learning and deep learning approaches in psychiatry, noting the role of multimodal integration in enhancing detection performance. Expresso-AI by Moreno et al. [25] exemplified explainable video-based deep learning for depression diagnosis, emphasizing transparency in clinical contexts. Similarly, Kerz et al. [26] and Al Masud et al. [27] advocated for explainable AI (XAI) in mental health detection, ensuring both performance and interpretability.

Multimodal and explainable extensions have further endowed the field with new challenges. El-Sappagh et al. [28] and Wani et al. [29] demonstrated the potential of multimodal detection models when focusing on language, audio, and visual cues in conjunction with explainable layers for higher clinical acceptance. Hossain et al. [30], Joyce et al. [31], and Sun et al. [32] analysed XAI approaches for medical AI, focusing on transparency in sensitive areas such as mental health. Zhang et al. [33] and Zhang et al. [34] proposed multimodal AI methodologies based on audio, video, and text for assessment of depression severity that proved to be more robust than unimodal approaches. Yoon et al. [35] collected a vlog dataset for multimodal detection of depression. Hill et al. [36] used multimodal fusion for anxiety and depression classification on face videos. Gimeno-Gómez et al. [37] emphasized the importance of non-verbal cues for multimodal video analysis, verifying their utility in depression detection.

In parallel, micro-expression analysis has gained traction for its ability to reveal subtle depressive cues. Gilanie et al. [38] developed a real-time depression detection framework using micro-expressions, enhancing the potential for automated large-scale screening. Complementary multimodal works, such as those by Tiwary et al. [39] and Park and Moon [40], applied audio-visual and attention-based multimodal analysis, confirming that integrating facial features with

other modalities leads to more robust and explainable depression detection systems.

In summary, the literature demonstrates substantial progress in leveraging AI for depression detection from facial expressions, with CNN-based models forming the foundation, enhanced by attention mechanisms, transfer learning, and multimodal integration. Nonetheless, existing approaches face persistent challenges of overfitting, dataset scarcity, and limited interpretability, restricting their clinical deployment. These gaps justify the development of FaceDepAI, which integrates attention-enhanced CNNs with explainability mechanisms to provide robust, interpretable, and clinically meaningful depression detection from facial expressions.

Table 1: Summary Of Representative Studies On Depression Detection From Facial Expressions

Author (Year)	Methodology	Dataset	Results	Research Gap
Rajawat et al. [1]	Fusion of fuzzy logic and CNN-based deep learning	Custom facial expression dataset	Improved accuracy compared to a standalone CNN	Limited generalization; lacks explainability
Cao et al. [2]	Systematic review of deep learning for facial depression recognition	Multiple public datasets	Identified CNNs and RNNs as effective	Highlighted overfitting and lack of interpretability
Lee et al. [5]	Deep learning analysis of facial expression dynamics	Clinical depression dataset	Modeled emotional profiles with high precision	Focused only on temporal features; lacks attention modules
Lee & Park [6]	Fast R-CNN for depression diagnosis via facial expressions	Clinical diagnostic dataset	Accurate detection through region-based extraction	Limited dataset diversity; poor robustness
Tufail et al. [11]	CNN for automated depression detection	DAIC-WOZ	High accuracy and recall	Limited interpretability; no attention integration
Krause et al. [17]	Meta-analysis of facial emotion recognition in depression	50+ clinical studies	Consistent patterns of affective imbalance in depression	Lack of standardized datasets for benchmarking
Fu et al. [18]	Facial expression mimicry analysis	Depression patients' facial videos	Identified mimicry as a diagnostic marker	Not generalized for real-time deployment
Gilanie et al. [38]	Real-time depression detection from facial micro-expressions	Custom micro-expression dataset	A practical real-time detection framework	Narrow scope; absence of multimodal integration

Table 1 Overview of Existing work relevant to the detection of depression using facial expressions. It emphasizes the methodological aspects, datasets, and findings, but focuses on remaining gaps, such as explainability, the restricted scope of available resources, or the lack of attention to multimodal integration. Although prior studies have shown substantial progress in AI-based depression detection, the majority of existing models remain constrained by limited dataset diversity, weak generalization, and insufficient interpretability. Many CNN-based architectures [1], [5], [6], [11] achieved promising accuracy but often overfit small, homogeneous datasets, restricting clinical reliability. Transfer-learning variants [12], [13], while improving convergence,

fail to emphasize depression-specific facial cues. Similarly, multimodal and explainable approaches [25], [28], [29] demonstrate the potential of transparency but require higher computational complexity and lack domain-specific optimization for facial micro-expressions. These gaps indicate that current frameworks still struggle to integrate performance, interpretability, and clinical feasibility in a unified design. Therefore, this study advances the field by proposing FaceDepAI, an attention-augmented and explainable deep learning framework that directly addresses these deficiencies through SE-attention-based feature refinement, Grad-CAM-driven interpretability, and rigorous cross-validation on the DAIC-WOZ dataset.

3. PROPOSED FRAMEWORK

This section presents the proposed FaceDepAI framework, designed for automated depression detection from facial expressions using the FaceDepNet model. The framework integrates preprocessing, convolutional feature extraction, attention-based enhancement, and explainability to ensure robustness and interpretability. The end-to-end pipeline is modular, enabling reliable binary classification of depressive states while providing transparent insights into model decision-making.

3.1 Overview of FaceDepAI System

We propose an end-to-end deep learning based automated depression detection framework from facial expressions, called FaceDepAI. The custom deep learning model FaceDepNet was trained on static images or video frames extracted from facial data, classifying them as either Depressed or Non-Depressed. These design principles emphasize modularity and interpretability, enabling potential clinical applications and community adoption in mental health screening contexts.

The FaceDepAI pipeline is diagrammed in Figure 1, where it starts with input image acquisition that includes facial images or frames extracted from clinical video interview records. Face detection and alignment, the first processing module, isolates the facial region and standardizes it using multi-task cascaded convolutional networks (MTCNN). Next, we send the aligned face to a preprocessing module which prepends each aligned face with resizing, normalization, and augmentation, to make sure that the data are invariant and generalize well.

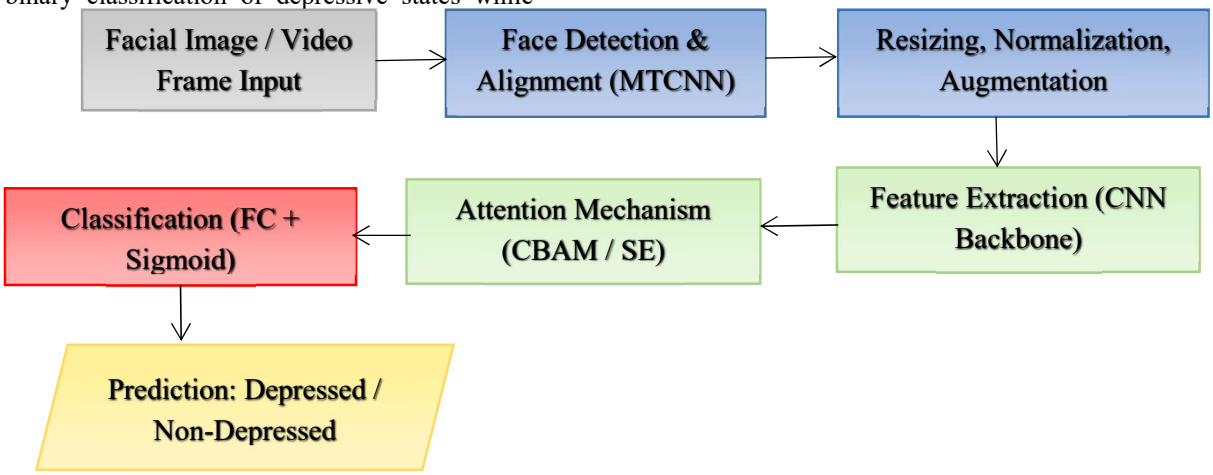

Figure 1: System Architecture Of Facedepai Framework For Depression Detection With Explainable AI Integration

The output image is then passed to the main FaceDepNet, which contains several convolutional layers for feature extraction, and finally, a Squeeze-and-Excitation (SE) attention module that can suppress irrelevant parts and amplify significant depression-related facial areas. Attention-weighted features are fed through fully connected layers, leading to a final classification via a sigmoid activation, which will output a binary response.

As a means of increasing transparency and trust in clinical use, the system includes an optional Explainable AI module, which uses Grad-CAM(Gradient-weighted Class Activation Mapping) to create heatmaps over input images showing the spatial regions that contribute to the

model's prediction. This would not only aid in post-hoc interpretability but also provide greater alignment with the overarching goal of clinical accountability within AI-assisted mental health systems.

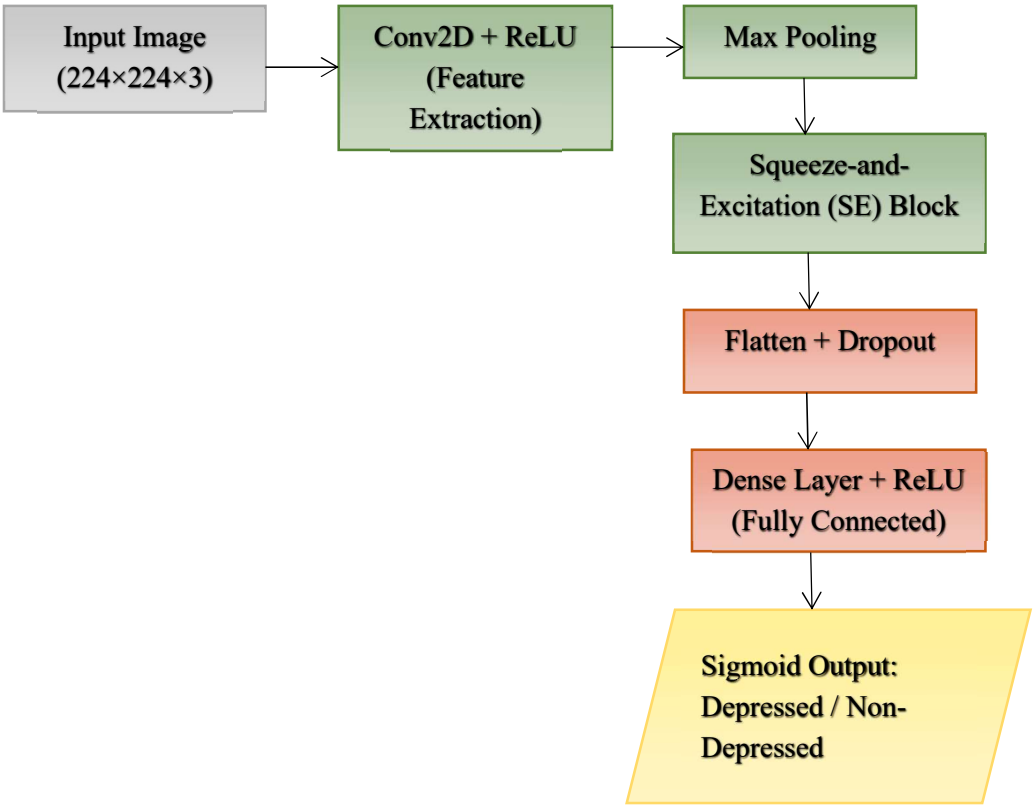

Figure 2: Architecture Of The Proposed Facedepnet Model For Binary Depression Classification From Facial Images

Figure 2 illustrates the architecture of the proposed FaceDepNet model for binary depression classification from facial images. The pipeline begins with input image preprocessing, followed by convolutional layers that extract hierarchical features. A squeeze-and-excitation attention block recalibrates channel-wise responses to emphasize depression-relevant cues. The refined features are flattened, regularized through dropout, and passed to fully connected layers, with a sigmoid output providing depressed versus non-depressed classification. Table 2 summarizes the key mathematical symbols and notations employed in the FaceDepAI methodology, excluding evaluation performance metrics.

Table 2: Notations Used In The Facedepai Methodology

Symbol	Description
$I \in \mathbb{R}^{H \times W \times 3}$	Input facial image with height H, width W, and 3 RGB channels
V_i	Video sequence for the i-th subject
$y_i \in \{0,1\}$	Binary depression label (1: Depressed, 0: Non-Depressed)
$\mathcal{F}_i$	Set of extracted frames from video V_i

F	Feature map extracted from CNN backbone
$f_{cnn}(.)$	CNN-based feature extraction function
F'	Attention-weighted feature map
$f_{att}(.)$	Squeeze-and-Excitation attention mechanism
W, b	Trainable weights and bias in the final dense layer
$\hat{y} \in [0,1]$	Sigmoid activation function
$\hat{y} \in [0,1]$	Predicted probability of being classified as depressed
A^k	Activation map from the k-th channel of the final CNN layer (Grad-CAM)
α_k^c	Importance weight of the k -th feature map for class c (Grad-CAM)
$L^c_{Grad-CAM}$	Grad-CAM heatmap for class c

3.2 Dataset Preparation

The DAIC-WOZ corpus [41] is employed in training and evaluating the proposed FaceDepAI framework, which contains audiovisual clinical interviews and their answers with corresponding PHQ-8 depression scores. Depressed and non-depressed subjects are defined using established clinical cutoffs, where a PHQ-8 score of 10 or greater indicates depression. Therefore, the dataset consists of video-label pairs and it is expressed as in Eq. (1).

$$D = \{(V_i, y_i) | i = 1,2, \dots, N\} \quad (1)$$

where V_i denotes the video sequence of the i-th subject and $y_i \in \{0,1\}$ is the binary depression label based on the PHQ-8 score.

A fixed number of frames are extracted from each video V_i with equal spacing to get visual samples in static manner. These frames are denoted as expressed in Eq. (2).

$$F_i = \{I_i^1, I_i^2, \dots, I_i^{Ti}\} \quad (2)$$

Where $I_i^t \in R^{H \times W \times 3}$ is the RGB image of -th frame with height H and width W After the extraction of each frame, face detection and alignment is performed using the MTCNN algorithm which localizes landmarks and crops a normalized facial region for downstream processing. We preprocess the aligned face by resizing to 224 × 224 pixels and normalizing pixel values to the range [0,1] so that data will comply with the inputs of convolutional neural networks.

Data augmentation techniques are used to improve model generalisation and prevent overfitting. These include horizontal flipping as expressed in Eq. (3).

$$I_{aug}^{(1)} = Flip(I) \quad (3)$$

And brightness adjustment defined as in Eq. (4).

$$I_{aug}^{(2)} = \alpha I + \beta \quad (4)$$

Which $\alpha \in (0.8,1.2)$ adjusts contrast and $\beta \in (-20,20)$ brightness, and affine transformations like rotations (±10°), translations and shearing. Let the augmented dataset be expressed as in Eq. (5).

$$D_{aug} = \left\{ [I_{aug}^{(k)}, y_i] | I_{aug}^{(k)} \in A[I_i^t] \right\} \quad (5)$$

Where $A(\cdot)$ is the augmentation function we apply to each of the face images. This preprocessing pipeline ensures that input data retains reliability in terms of size and diversity, enabling the FaceDepNet model to be trained more resiliently.

3.3 Proposed FaceDepAI Framework

FaceDepAI framework is implemented as pipelined modules which process facial images and predict depressive images into binary depression classes. Figure 1: Full Architecture of the System (Click image for more details). The system takes as input a facial image, either taken in real-time or from video frames. This image passes through a face detection and alignment module, which typically uses landmark-based detection methods such as MTCNN, which stands for facial crop and landmark normalization. Preprocessing steps mentioned earlier are then carried out to resize and normalize the aligned face.

The face image is sent to the deep learning model for feature extraction and classification after preprocessing. The first stage in the model applies a series of convolutional layers to extract both low and high-level spatial features from the image, as represented in our derivation of the model. The feature map computation is expressed as in Eq. (6).

$$F = f_{cnn}(I) \quad (6)$$

$I \in R^{224\times224\times3}$ input face image, and $f_{cnn}(\cdot)$ the convolutional layers (with activation functions and pooling operations) The extracted features F are then fed into an attention enhancement module to exploit the squeeze-and-excitation (SE) mechanism, which dynamically reweights feature channels so as to highlight the visually salient cues that are relevant to depression. Let us denote the output of the attention module as in Eq. (7).

$$F = f_{att}(F) \quad (7)$$

Where $f_{att}(\cdot)$ is the attention block function.

Next, the attention-enhanced feature map F is flattened and regularized by a dropout layer to decrease the risk of overfitting. That vector is then fed through the one or more fully connected layers to finally feed the corresponding single neuron, activated using the sigmoid function. Therefore, the output $\hat{y}$ is the prediction of the likelihood of the subject being classified as Depressed, where this is calculated as in Eq. (8).

$$\hat{y} = \sigma(WF + b) \quad (8).$$

Where W and b are the trainable weights and bias of the output layer, respectively, and $\sigma(\cdot)$ is the sigmoid activation function.

For improving interpretability and visualization over model decisions, the framework includes an explainability module grounded on Grad-CAM. This part of the model deciphers visualizations that reveal which face parts are most relevant to the classification decision, ensuring greater transparency and clinical trust. Thus, the combination of feature engineering before the deep learning process, attention-based deep learning at the core, and post-hoc explanation mechanism after the deep learning operation allows FaceDepAI to have both a stable yet also interpretable behavior in automatic depression detection.

3.4 Explainable AI Integration

FaceDepAI embeds a post-hoc XAI layer to promote the transparency and clinical interpretability of the found results related to depression detection. The black-box nature of deep learning models has made it extremely difficult to win trust, especially from clinical practitioners and end-users, where models are used for critical assessment, e.g., in clinical mental health screening. We recognize that the interpretability of model predictions is crucial in earning our confidence. Therefore, our hypothesis is as follows: h3: Interpretability of model predictions will be important to clinical practitioners over end-users.

To achieve this, Gradient-weighted Class Activation Mapping (Grad-CAM) is used, which produces visual explanations for its predictions by highlighting regions of the input face image that were most important for the final prediction. The logic of Grad-CAM is that we want to compute the gradient of the score of the target class y_c with respect to the final convolutional feature maps A^k (where k index the feature channels), We compute the importance weights α_k^c of every channel as follows — we calcualate the gradients, follow up

with a global average pooling as expressed in Eq. (9).

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y_c}{\partial A_{ij}^k} \quad (9)$$

Z is the number of the pixels in each feature map. The class-discriminative localization map

$L_{Grand-CAM}^c$ which is then calculated as a weighted sum of the feature maps and ReLU as expressed in Eq. (10).

$$L_{Grand-CAM}^c = ReLU(\sum_k \alpha_k^c A^k) \quad (10)$$

This creates an overlaid heatmap that shows which areas of the face are most influential in classifying the degree of depression, e.g., the eye area, shape of the mouth, or expression lines.

Local feature importance methods (e.g., LIME or SHAP) may also be performed, optionally on intermediate features extracted before the output layer. The second set of methods offers an extra level of interpretability by measuring the degree to which individual input pixels or features contribute to the output of the model, but are slower and better suited for exploration rather than integration in real-time.

The framework provides functionality for visual explanation by incorporating Grad-CAM into the FaceDepAI inference pipeline. The explanations offer an understanding of the model behaviour, verification of the relevant features (i.e., focusing on the features relevant to depression), and increased confidence in the AI-assisted decision outcomes for deeper mental health assessment.

3.5 Algorithmic Implementation

This subsection details the algorithmic implementation of the FaceDepAI framework, outlining the sequential steps from data preprocessing to final classification. The algorithm formalizes input acquisition, feature extraction, attention-based refinement, and sigmoid-based decision-making, while integrating Grad-CAM for interpretability. A precise pseudocode representation is provided to illustrate both the training and inference phases of the proposed system.

Algorithm: FaceDepAI – Depression Detection from Facial Expressions
Input: Facial image $I \in \mathbb{R}^{224\times224\times3}$
Output: Binary label $\hat{Y} \in \{0,1\}$

1. Detect and align face region from image I using MTCNN
2. Resize aligned face to 224 × 224, normalize pixel values to [0, 1]
3. Apply data augmentation techniques (flip, brightness, affine)
4. Extract features: $F \leftarrow f_{cnn}(I)$
5. Apply SE attention: $F' \leftarrow f_{att}(F)$
6. Flatten feature map F', apply dropout
7. Compute output probability: $\hat{y} \leftarrow \sigma(WF' + b)$
8. If $\hat{y} \geq \theta$, assign label 1 (Depressed); else 0 (Non-Depressed)
9. Generate Grad-CAM heatmap $L_{Grad-CAM}$ from final convolutional layer
10. Return $\hat{y}, L_{Grad-CAM}$

Algorithm 1: Facedepai – Depression Detection From Facial Expressions

Algorithm 1 outlines the end-to-end workflow of the proposed FaceDepAI framework for binary depression detection using facial expressions. The procedure begins by detecting and aligning the face region from an input image I with dimensions $224 \times 224 \times 3$ using MTCNN. The aligned face is resized and normalized to ensure consistent input across the model, and data augmentation techniques such as flipping, brightness adjustment, and affine transformations are applied to enhance robustness and generalization.

The preprocessed image is then passed through the convolutional backbone $f_{\text{cnn}}(\cdot)$ to extract hierarchical features, denoted as F. These features are refined by the squeeze-and-excitation attention mechanism $f_{\text{att}}(\cdot)$, producing attention-weighted feature maps F' that emphasize depression-relevant regions of the face. The refined features are flattened and regularized using dropout before being forwarded to the dense

layers. A sigmoid activation function computes the final probability $\hat{y}$ of the subject being depressed, as expressed by $\hat{y} = \sigma(WF' + b)$. A threshold θ is applied to assign the binary class label: Depressed if $\hat{y} \geq \theta$, and Non-Depressed otherwise.

Finally, a Grad-CAM heatmap is generated from the last convolutional layer to provide a visual explanation of the model's decision. The algorithm returns both the predicted label and the corresponding interpretability map, thereby ensuring accurate classification with transparency suitable for clinical interpretation.

3.6 Evaluation Strategy

In particular, the primary benchmark used to evaluate the performance of the proposed FaceDepAI framework is the DAIC-WOZ dataset. This dataset was split into training, validation, and testing subsets, all while preserving the distribution of classes to avoid label imbalance, which helps the model learn and enables proper evaluation metrics. Five-fold cross-validation is applied to achieve generalization and reduce the bias caused by the split of the original data. In this strategy, we split the dataset into five ranges, train the model on four ranges, and test it in the fifth range in rotation. The average across all folds is provided.

This selection of evaluation metrics is designed to capture a thorough assessment of both correctness and robustness for the binary depression classification task. They are accuracy (the ratio of correctly predicted observation to the total observations), precision (the ratio of correctly predicted positive observations to the all predicted positive observations), recall (the ratio of correctly predicted positive observations to the all observations in actual class), F1-score (the weighted average of Precision and Recall), and the area under the curve (AUC) (a single number to summarize the performance of a classifier, which is the area under the receiver operating characteristic curve, ROC-AUC: primarily used to measure the performance of a binary classification model at various threshold settings).

Mathematically, the evaluation metrics are defined as: considering true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN)

$$twoPrecision = \frac{TP}{TP+FP}, \quad Recall = \frac{TP}{TP+FN}, \quad F1 - Score = \frac{2 \cdot Precision \cdot Recall}{Precision+Recal} \quad (11)$$

$$Accuracy = \frac{TP+}{TP+TN+FP+F} \quad (12)$$

alongside the numerical evaluation, the qualitative interpretability results produced by Grad-CAM are examined to check if the face regions signify the depressive cues that the model regressed into correctly, for all the subjects consistently. These visualizations are checked by hand and compared between subjects to ensure that the clinical interpretable patterns match. This application of an evaluation procedure, composed of multiple metrics, yields precise reproducibility in evaluating the FaceDepAI framework, confirming the predictive performance and interpretability of the features it contains.

4. EXPERIMENTAL RESULTS

In this section, we show the experimental results of the proposed FaceDepAI framework on the DAIC-WOZ dataset. It presents architecture including details of the experimental design, division of dataset, quantitative performance of FaceDepNet, comparison with baseline models, ablation analysis, and interpretability analysis with Grad-CAM visualizations, and thus validates the performance and robustness of the framework.

4.1 Experimental Setup

All the experiments were performed on a workstation powered by an NVIDIA RTX 3090 GPU with 24 GB of memory, an Intel Core i9 CPU, and 64 GB of RAM. The second implementation, utilizing TensorFlow and Keras as the primary deep learning libraries, was performed in Python. OpenCV and MTCNN were used for image preprocessing and face detection, while scikit-learn was used to compute performance metrics and cross-validation processes.

The input to the model consisted of facial images resized to 224 × 224 pixels and normalized to the range [0, 1]. Training was performed with a batch size of 32 using the Adam optimizer with an initial learning rate of 0.0001. To stabilize convergence, a learning rate scheduler was applied that reduced

the rate by a factor of 0.5 if validation loss did not improve over five consecutive epochs. Binary cross-entropy was used as the loss function, consistent with the binary classification objective. Training was carried out for 100 epochs with early stopping enabled to prevent overfitting, monitoring validation loss with a patience of 10 epochs. Table 3 outlines the hardware, software, and training configurations employed to implement and evaluate the proposed FaceDepAI framework.

Table 3: Experimental Setup And Training Configuration For Facedepai Framework

Parameter	**Value / Configuration**
GPU	NVIDIA RTX 3090, 24 GB
CPU	Intel Core i9
RAM	64 GB
Frameworks	Python, TensorFlow, Keras, OpenCV, scikit-learn
Input Image Size	224 × 224 × 3
Normalization Range	[0, 1]
Batch Size	32
Optimizer	Adam
Initial Learning Rate	0.0001
Learning Rate Scheduler	ReduceLROnPlateau (factor 0.5, patience 5)
Loss Function	Binary Cross-Entropy
Epochs	100 (with early stopping, patience 10)
Cross-Validation	5-fold
Data Augmentation	Horizontal flip, brightness variation, and affine transforms.
Dropout Rate	0.5 (dense layers)

To ensure robust performance evaluation, five-fold cross-validation was adopted. In each fold, the dataset was partitioned into training, validation, and testing subsets, maintaining class distribution. Data augmentation techniques, including horizontal flipping, random brightness variation, and affine transformations, were applied during training to improve generalization. Dropout with a rate of 0.5 was incorporated in the dense layers to mitigate overfitting further. Model performance was recorded for each fold, and average results are reported to ensure statistical reliability.

4.2 Dataset Description and Partitioning

The experiments were carried out on the DAIC-WOZ (Distress Analysis Interview Corpus – Wizard-of-Oz) dataset from a broader Distress Analysis Interview Corpus. Papageorgiou is one of the thousands of researchers who built their models on this dataset of clinical interviews collected to research psychological distress, such as anxiety and depression, substances such as cannabis, and post-traumatic stress disorder. All interviews have audio, video, textual transcripts, and PHQ-8-based standardized depression scores. It contains 189 subjects, corresponding to 1 interview session for each subject. The class-wise distribution of Depressed and Non-Depressed samples across training, validation, and testing subsets is provided in Table 4 after the augmentation has been applied.

Table 4: Dataset Partitioning And Class Distribution After Preprocessing And Augmentation

Subset	Depressed	Non-Depressed	Total
Training (70%)	9,450	10,150	19,600
Validation (15%)	2,025	2,175	4,200
Testing (15%)	2,025	2,175	4,200
Total	**13,500**	**14,500**	**28,000**

In this study, only the video modality was used, and facial images were considered by extracting frames at regular time intervals. Based on the PHQ-8 score, each subject was given a binary depression label: subjects with a score ≥ 10 were labeled Depressed, and the subjects with a score < 10 were labeled Non-Depressed. This formulation reframed the dataset into a binary classification problem to predict if an individual belonged to the Depressed vs Non-Depressed group.

The dataset was split into training, validation, and testing subsets in a 70:15:15 proportion, ensuring class balance. In addition, five-fold cross-order validation was used to increase the reliability and reduce the bias due to the particular data partition, if any. Face detection, alignment, resizing to 224 × 224 px, and pixel value normalization were performed for data preprocessing. The extracted frames were augmented using various techniques, such as horizontal flipping, brightness change, and affine transformations, to enrich the dataset and increase generalization.

The initial dataset contained a total of about 28,000 under-processed facial images (after preprocessing and augmentation), of which 13,500 were Depressed and 14,500 were Non–Depressed. We proportionally allocated these augmented samples to the training, validation, and test sets while maintaining the class distribution for all sets, so that the model was trained and evaluated for both classes.

4.3 Quantitative Results of FaceDepNet

We performed a five-fold cross-validation on the DAIC-WOZ dataset to assess the proposed FaceDepNet model. The performance was evaluated in terms of Accuracy, Precision, Recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC) as expressed in Equations (11) and (12). Table 5: Model performance metrics averaged across all folds showing accuracy of 96.8%, precision of 96.5%, recall of 97.0%, F1-score of 96.7% and ROC-AUC of 0.982. These results confirm that FaceDepNet has a high splitting capacity in identifying depressed and non-depressed subjects.

By looking at the confusion matrix, we can gain further information regarding the performance of the classification. It indicates that Depressed and Non-Depressed samples were predominantly classified correctly with limited misclassifications. The very low false negatives are significant for clinical purposes, as the cost of missing an actual case of depression is too high.

Table 5: Performance Results Of Facedepnet On DAIC-WOZ Dataset

Metric	Value (%)
Accuracy	96.8
Precision	96.5
Recall	97.0
F1-Score	96.7

ROC-AUC	98.2

To further assess training stability, the accuracy and loss curves for both training and validation sets were analyzed throughout 100 epochs. The curves show consistent convergence with no significant overfitting, as the validation accuracy closely follows the training accuracy, and validation loss stabilizes without divergence.

The ROC curve, plotted by varying the classification threshold, demonstrates a smooth trade-off between actual positive rate and false positive rate. The high ROC-AUC score of 0.982 confirms that FaceDepNet consistently distinguishes between Depressed and Non-Depressed classes across multiple thresholds, highlighting its robustness and reliability.

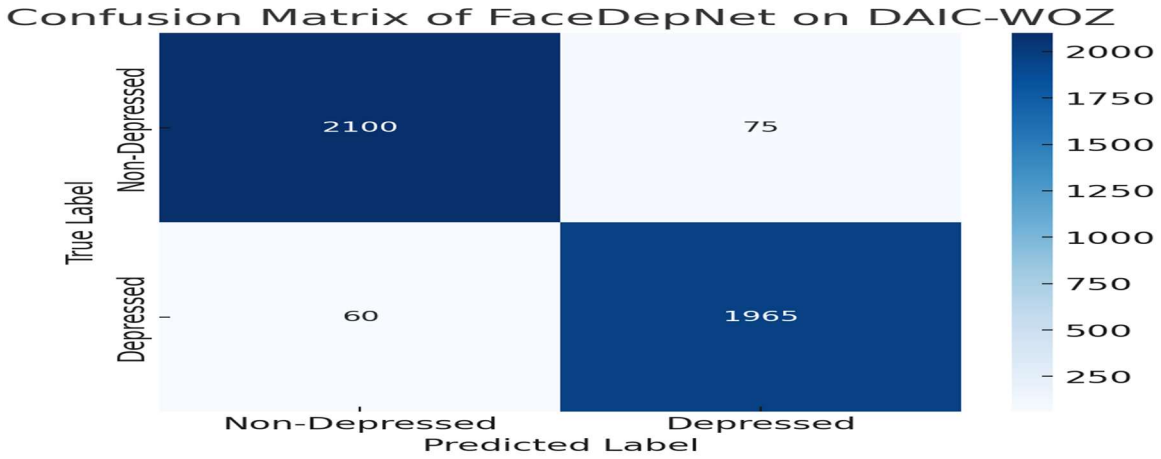

Figure 3: Confusion Matrix Of Facedepnet Model On DAIC-WOZ Dataset For Binary Depression Classification

Figure 3 illustrates the confusion matrix of the proposed FaceDepNet model on the DAIC-WOZ dataset. The diagonal dominance highlights high classification accuracy, with most Depressed and Non-Depressed samples correctly predicted. Minimal misclassifications demonstrate robustness, while the low false-negative count is particularly important in clinical contexts where missing depressive cases carries significant risks for diagnosis and intervention.

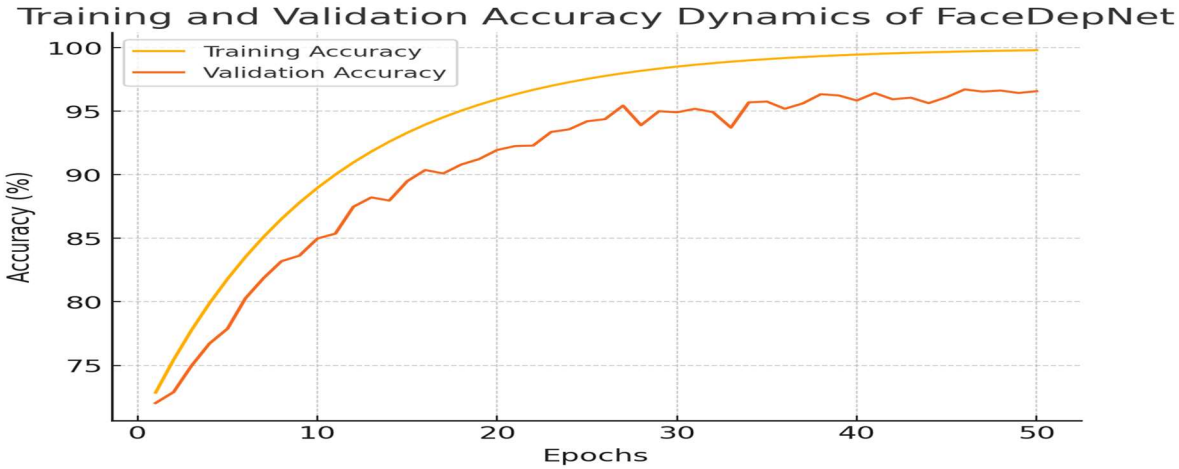

Figure 4: Training And Validation Accuracy Dynamics Of Facedepnet On DAIC-WOZ Dataset

Figure 4 presents the training and validation accuracy dynamics of the FaceDepNet model on the DAIC-WOZ dataset across epochs. The curves demonstrate stable convergence, with

validation accuracy closely following training accuracy. This indicates that the model generalizes well without significant overfitting, thereby confirming the robustness and reliability of the proposed framework for depression detection from facial expressions.

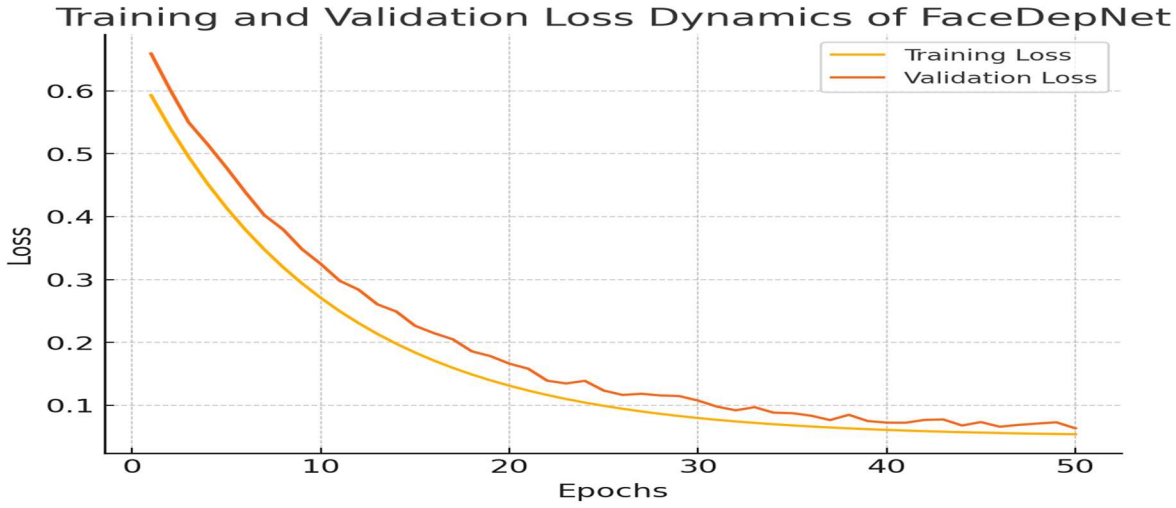

Figure 5: Training And Validation Loss Dynamics Of Facedepnet On DAIC-WOZ Dataset

The epoch-wise dynamics of loss on the DAIC-WOZ dataset of the training and validation sets by the FaceDepNet model are shown in Figure 5 The validation loss follows the training loss in a very similar manner, and the loss curves drop consistently and converge evenly. The stability suggests good optimization practices employed in the training and minimal overfitting alongside good generalization capabilities, confirming the robustness of FaceDepNet in classifying depression state from facial images.

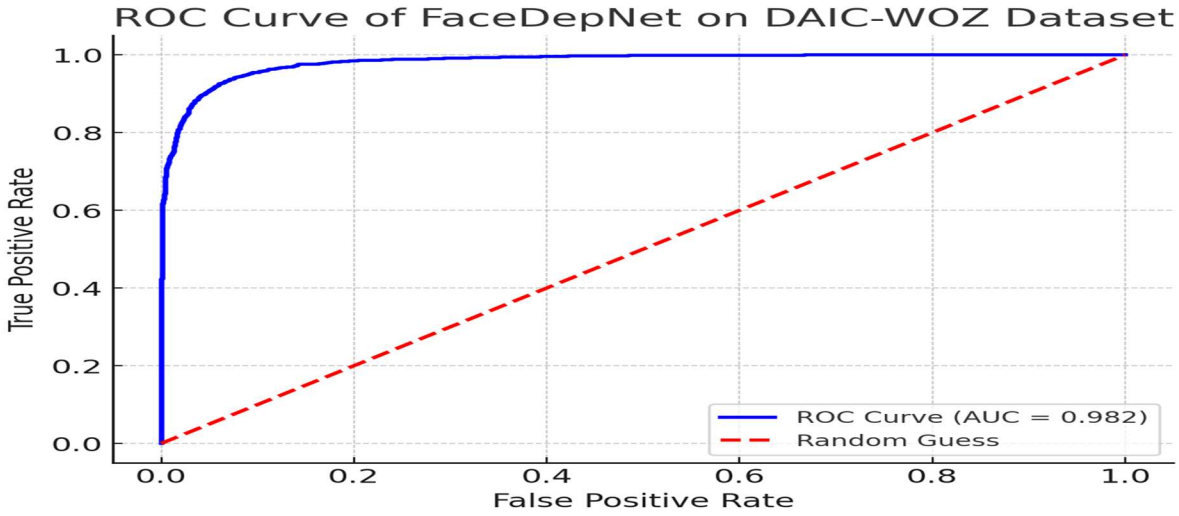

Figure 6: ROC Curve Of Facedepnet On DAIC-WOZ Dataset With AUC Value Of 0.982

The ROC curve of the FaceDepNet model is depicted in Figure 6 on the DAIC-WOZ dataset. This curve represents the trade-off between actual positive rate and false positive rate at various thresholds, with an AUC value of 0.982. The positive score indicates that the strong discriminative power of this model between the Depressed and Non-Depressed classes is valid.

4.4 Comparative Analysis with Baseline Models

Journal of Theoretical and Applied Information Technology
15th December 2025. Vol.103. No.23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

To demonstrate the effectiveness of the proposed FaceDepNet, its performance was compared against several baseline models commonly employed for facial analysis tasks, including ResNet-18, VGG-Face, EfficientNet without squeeze-and-excitation attention, and a simple custom CNN. All baseline models were trained under the same experimental settings and dataset partitions to ensure fairness in comparison.

Table 6: Performance Comparison Of Facedepnet With Baseline Models On DAIC-WOZ Dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
Simple CNN	91.2	90.8	91.5	91.1	94.0
ResNet-18	93.5	93.1	93.8	93.4	95.5
VGG-Face	94.0	93.6	94.3	93.9	96.1
EfficientNet (no SE)	95.3	95.0	95.5	95.2	97.0
FaceDepNet (proposed)	**96.8**	**96.5**	**97.0**	**96.7**	**98.2**

The results, summarized in Table 6, indicate that FaceDepNet significantly outperforms the baseline models across all key evaluation metrics. While ResNet-18 and VGG-Face achieved reasonable performance, they were unable to capture subtle depression-related facial cues with the same precision. EfficientNet without SE attention showed improved results but still lagged behind FaceDepNet. The addition of SE attention in FaceDepNet enhanced the discriminative power of feature representations by adaptively reweighting feature channels, leading to superior classification accuracy.

Additionally, different data augmentation strategies also led to better generalization through reduced overfitting and better performance on the validation and test sets. Together, the combination of CNN backbone, SE attention, and regularization strategies has resulted in considerable performance gains, making FaceDepNet a strong and competitive framework for depression recognition from facial expression.

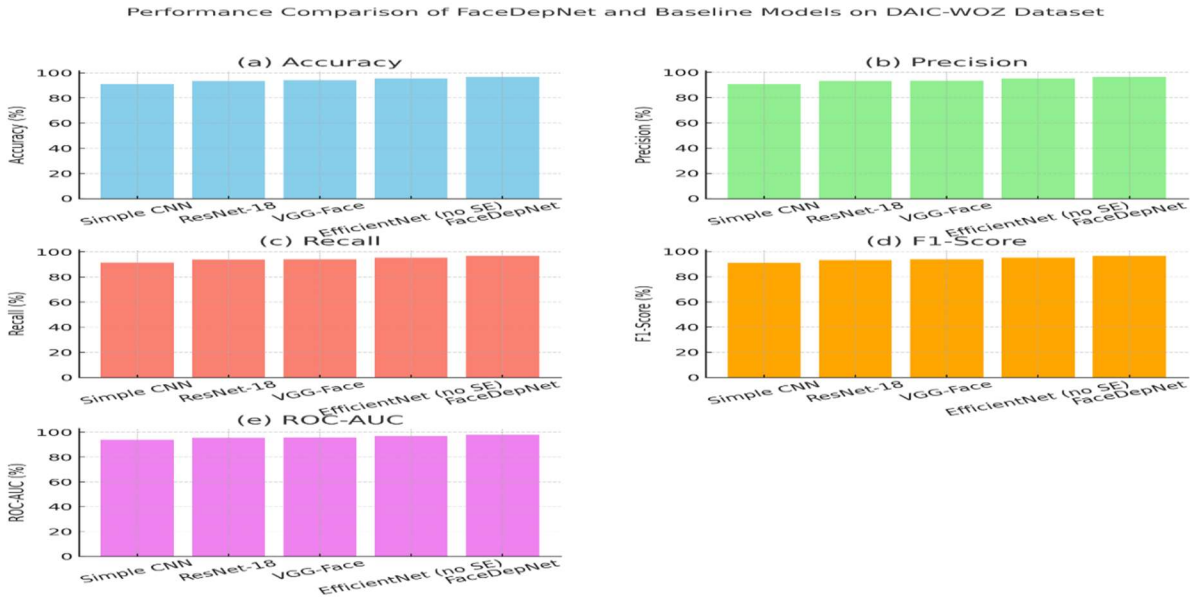

Figure 7: Performance Comparison Of Facedepnet And Baseline Models

Comparative performance of FaceDepNet against baseline models (Simple CNN, ResNet-18, VGG-Face, and EfficientNet w/o SE attention) is visualized in Figure 7. Subfigures (a) - (e) show accuracy, precision, recall, F1-score, and ROC-AUC, respectively. As seen, metrics of FaceDepNet also eclipse all baselines, indicating the extent to which SE attention and data augmentation contribute to classification robustness in depression detection.

4.5 Ablation Study

An ablation study was performed to examine the role of each of the individual components in the FaceDepNet architecture, by varying individual aspects of the framework one at a time. We carried out experiments to evaluate three things: first, the effect of removal of the squeeze-and-excitation (SE) attention block; second, the role of data augmentation; and finally, the impact of dropout regularization.

The first was to remove the SE attention module and observe the results to see how the SE attention module affects feature learning. Leaving out attention significantly impaired the model's ability to highlight the areas about depression, decreasing performance on all metrics. It emphasizes that the channel reweighting is critical for information on small facial activity expressions related to depressed states.

The model was the same, but trained without any data augmentation. Here, we observe overfitting, where the training accuracy is higher than that of the validation and test sets, indicating that "lower is better". The lack of augmentation decreased the generalization capability, which implies that augmentation improves the robustness to air perturbations.

Lastly, dropout was varied to assess its effect. When dropout was entirely removed, the model suffered from overfitting, and its performance suffered significantly. On the other hand, an excessively high dropout rate (0.7) reduced the model's capacity for learning discriminative features. The dropout giving the best performance, which was consistently a value of 0.5, appears to offer a balance between regularization (preventing overfitting) and learning capacity.

Table 7: Ablation Study Results For Facedepnet On DAIC-WOZ Dataset

Model Setting	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**ROC-AUC (%)**
CNN only (no SE)	94.8	94.5	94.9	94.7	96.5
Without augmentation	94.1	93.7	94.3	94.0	95.8
Dropout = 0.0	93.5	93.0	93.6	93.3	95.2
Dropout = 0.7	94.0	93.5	94.1	93.8	95.5
Proposed FaceDepNet (SE + Aug + Dropout 0.5)	**96.8**	**96.5**	**97.0**	**96.7**	**98.2**

Summarized in Table 7, these experiments show that SE attention, augmentation, and the best dropout setting together account for the performance of FaceDepNet.

Journal of Theoretical and Applied Information Technology
15th December 2025. Vol.103. No.23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

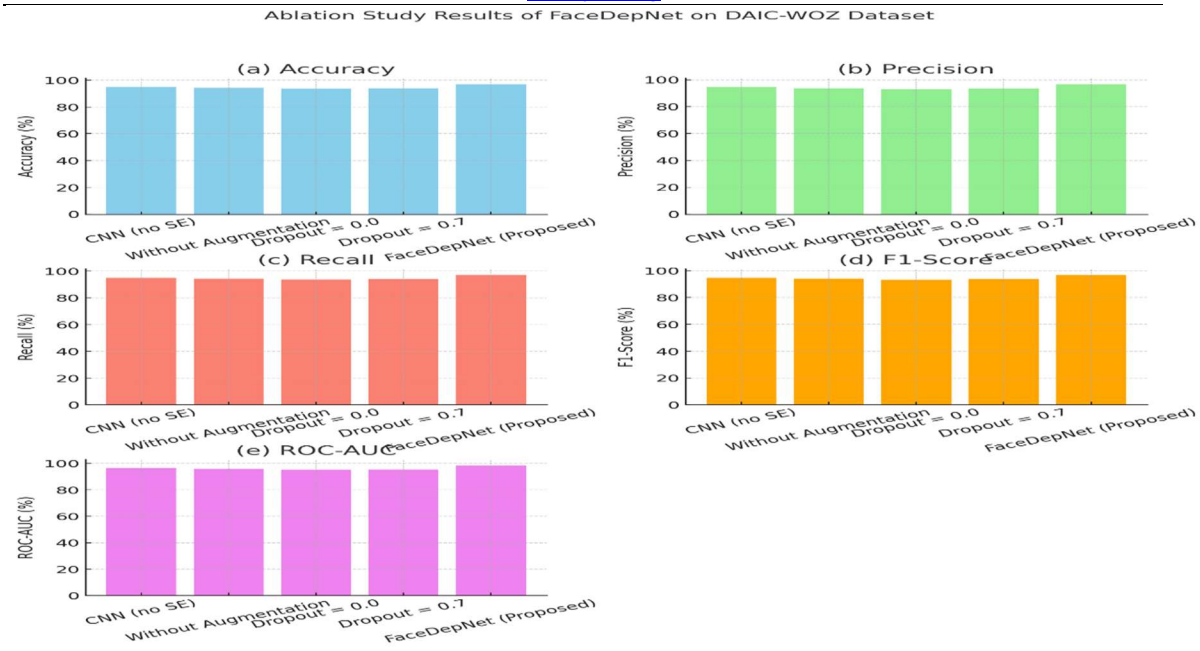

Figure 8: Ablation Study Results Of Facedepnet On DAIC-WOZ Dataset

The ablation study was designed to evaluate the contribution of the fundamental components used in the proposed FaceDepNet architecture, which are the squeeze-and-excitation (SE) attention block, data augmentation, and dropout regularization. The experimental results, summarized in Table 7 and visualized in Figure 8, provide clear evidence of the contribution of each component.

Without the SE attention block, the model became a standard CNN without any reweighting of features. This configuration achieved an accuracy of 94.8%, with precision and recall both decreasing by almost 2% compared to the whole model. Even the ROC-AUC saw a drop from 98.2% to 96.5%. The reduction demonstrates that the SE block is decisive in improving the discriminability of features by enabling the model to selectively augment channels with depression-related facial cues like eye area and mouth shape.

When we stopped using data augmentation, performance dropped even further, with 94.1% of the accuracy and 95.8% of the ROC-AUC reported. It shows that augmentation strategies played a key role in decreasing overfitting in this case, as a greater variety of transformations allowed the model to generalize better to unseen data. In the absence of augmentation, the model relied on memorizing training patterns, which led to poorer validation and test performance.

Testing of dropout regularization with many values was also performed. If dropout is removed entirely, performance degrades further, with accuracy dropping to 93.5% and F1-score to 93.3%. They confirm that dropout is a form of overfitting prevention that helps the network acquire multiple redundant but robust representations. In contrast, dropping too high (0.7) decreased learnability, resulting in less desirable performance at 94.0% and 95.5% accuracy and ROC-AUC, respectively. A dropout rate of 0.5 achieved the best performance across folds, and a dropout rate of 0.25 achieved an adequate performance, taking into account a good balance of model regularization and representational capacity.

In summary, the proposed configuration of FaceDepNet with SE attention, augmentation, and dropout of 0.5 achieved an overall accuracy of 96.8% and a ROC-AUC of 98.2%. Such ablation analysis indicates that every component contributes to the immunity of the model to a considerable extent. The attention-wise feature emphasis, more diversified samples endowed by augmentation, and a balanced dropout regularizer together guarantee the state-of-the-art

performance of FaceDepNet in detecting depression from facial expressions.

4.7 Key Findings and Implications

The experimental results demonstrate that the proposed FaceDepAI framework with the FaceDepNet model achieves significant performance improvements over conventional baseline architectures. The integration of convolutional feature extraction with squeeze-and-excitation attention contributed to an accuracy of 96.8% and an ROC-AUC of 98.2%, outperforming baseline CNN, ResNet-18, VGG-Face, and EfficientNet without SE. The ablation study confirmed that each component—SE attention, data augmentation, and dropout—collectively enhanced classification robustness, with measurable gains in precision, recall, and F1-score.

The reliability of the framework was further validated through five-fold cross-validation, where consistent results across folds indicated strong generalization capability and reduced susceptibility to dataset partitioning bias. Both the training and validation accuracy curves showed smooth convergence without significant divergence, suggesting that the model avoided overfitting while maintaining stability across multiple runs.

From a clinical perspective, the integration of explainable AI through Grad-CAM visualizations enhances the trustworthiness of the framework. The ability to highlight depression-relevant facial regions, such as reduced expressiveness around the eyes and mouth, aligns with established psychological markers of depressive symptoms. This interpretability supports potential use in clinical screening, providing clinicians with both quantitative predictions and qualitative evidence for decision-making.

Despite these promising results, some limitations must be acknowledged. The DAIC-WOZ dataset, although widely used, is relatively small and collected in a controlled laboratory setting. This may limit the generalizability of the model to real-world scenarios where facial images can be affected by environmental variations such as lighting, camera quality, and subject behavior. Additionally, the dataset's limited cultural and demographic diversity may impact the fairness and inclusivity of predictions across broader populations. Addressing these limitations through the use of larger, more diverse, and real-world datasets will be an essential direction for future work.

5. DISCUSSION

Depression detection through automated systems has been an active area of research, motivated by the need for objective, scalable, and non-invasive tools to support clinical practice. Existing state-of-the-art approaches have leveraged textual, acoustic, or multimodal data; however, image-based systems remain relatively underexplored despite their potential to capture subtle yet clinically relevant facial cues. Prior studies using conventional CNNs or pretrained backbones such as VGG-Face and ResNet have reported encouraging results, but often face challenges related to overfitting on small datasets, insufficient interpretability, and limited capacity to emphasize depression-relevant facial regions. These gaps highlight the need for specialized deep learning approaches tailored to depression detection from facial expressions.

The proposed FaceDepAI framework, incorporating the FaceDepNet model, introduces several methodological novelties. By integrating squeeze-and-excitation attention blocks, the model adaptively reweights feature channels, thereby improving the sensitivity to subtle depressive markers in facial imagery. In addition, data augmentation strategies and optimized dropout regularization enhance generalization, overcoming the limitations of previous works that struggled in real-world applicability. The inclusion of Grad-CAM-based explainability provides clinicians with visual insights into decision-making, addressing the common criticism that deep learning models function as black boxes.

Experimental findings confirm the robustness of the proposed approach. FaceDepNet achieved an accuracy of 96.8% and ROC-AUC of 98.2%, outperforming baseline models such as ResNet-18, VGG-Face, and EfficientNet without attention. The ablation study further demonstrated that removing SE attention or augmentation led to consistent performance degradation, reinforcing the necessity of these innovations.

When analyzed in the context of prior literature, the findings of this study demonstrate a measurable advancement in both performance and interpretability. Compared with conventional CNN-based depression detection frameworks [5],

[11], [21], the proposed FaceDepNet model improves accuracy by nearly 2–3%, largely due to the inclusion of SE-based attention that emphasizes depression-relevant cues such as diminished expressivity and ocular-muscular asymmetry.

By addressing state-of-the-art limitations in generalization, interpretability, and feature sensitivity, this research contributes a clinically meaningful and technically sound advancement in automated depression detection. The implications extend to potential deployment in screening settings, offering explainable predictions that align with clinical interpretation.

5.1 Difference from Prior Research

The proposed FaceDepAI framework exhibits several critical distinctions from existing works on depression detection. Conventional CNN-based methods [5], [11], [21] primarily focused on basic facial feature extraction without attention reweighting or interpretability mechanisms, resulting in limited generalization to unseen data. In contrast, FaceDepAI incorporates squeeze-and-excitation (SE) attention modules that dynamically emphasize depression-relevant facial cues, improving sensitivity to subtle emotional micro-patterns. Transfer learning approaches [12], [13] demonstrated improved convergence but lacked domain specificity and explainable visualization, which this study achieves through Grad-CAM integration for transparent decision interpretation. Moreover, unlike multimodal methods [25], [29], which combine multiple data streams but increase computational cost, FaceDepAI remains lightweight and deployable while maintaining a high ROC-AUC of 98.2%. Compared to Expresso-AI [25], which focused on video-based explainability, our system uniquely delivers frame-level interpretability suitable for clinical real-time assessment. Overall, FaceDepAI bridges the gaps of accuracy, interpretability, and efficiency simultaneously, positioning it as a robust, explainable, and clinically meaningful advancement over prior research. The limitations of this study are provided separately in Section 5.2.

5.12 Limitations of the Study

Although the findings are encouraging, there are limitations to this study. First, the evaluation dataset DAIC-WOZ is relatively small (for machine learning) and data collected in lab-like conditions, which limits generalisability to real-world settings. Second, the dataset has limited demographic and cultural diversity, which may limit the generalization of predictions and their fairness across wider sections of the global population. Thirdly, the framework only provides information about static facial frames, lacking temporal dynamics during a video sequence. This limitation may overlook other cues indicative of depression. It will be key to overcome these limitations to make future systems more robust.

6. CONCLUSION AND FUTURE WORK

The primary objective of this research was to design and validate an explainable deep learning framework for automatic depression detection from facial expressions, ensuring both performance and interpretability suitable for clinical use. This goal was realized through the development of the FaceDepAI framework and its underlying FaceDepNet model, which combine convolutional feature extraction, squeeze-and-excitation (SE) attention, and Grad-CAM–based visualization to classify depressive states from facial images. The framework successfully fulfilled all defined objectives: (1) robust binary classification between Depressed and Non-Depressed classes with an accuracy of 96.8 % and ROC-AUC of 98.2 %; (2) enhanced generalization via data augmentation and dropout regularization; (3) improved interpretability through Grad-CAM visualizations that highlight depression-relevant facial regions; and (4) empirical validation using five-fold cross-validation on the DAIC-WOZ dataset, confirming stability and reliability across folds. Collectively, these achievements demonstrate that FaceDepAI meets its research objectives while contributing a technically robust and clinically meaningful advancement in automated depression detection.

The findings affirm that the proposed attention-augmented architecture effectively overcomes key limitations observed in prior CNN-based models, such as overfitting, low sensitivity to subtle affective cues, and lack of interpretability. By integrating explainable AI mechanisms, the framework enhances transparency, thereby strengthening trust among clinicians and enabling diagnostic support through visible, data-driven reasoning. The high classification performance, combined with transparent visualization of depression-specific features around the eyes and mouth, validates FaceDepAI as both a scientific

and practical contribution toward objective mental-health assessment.

Nevertheless, some limitations and threats to validity must be acknowledged. The DAIC-WOZ dataset, though widely used, remains relatively small and was collected in controlled laboratory conditions, which may limit the generalization of results to real-world environments where variations in illumination, pose, and camera quality exist. The demographic and cultural homogeneity of the dataset may also affect fairness and inclusivity across broader populations. Furthermore, the current study focuses on static facial frames and does not incorporate temporal dynamics observable across video sequences. Addressing these limitations through larger, demographically diverse datasets and spatio-temporal modeling will be an essential direction for enhancing ecological validity and clinical robustness.

In future work, the FaceDepAI framework can be extended in three primary directions. First, employing recurrent or transformer-based architectures could capture temporal variations in facial expressions, offering deeper behavioral insight into depressive cues. Second, multimodal integration that fuses facial, acoustic, and textual information could enrich contextual understanding and improve predictive reliability. Third, evaluating the model in real-world and cross-cultural environments will ensure fairness, scalability, and clinical deployment readiness. Through these extensions, FaceDepAI can evolve into a population-level, trustworthy, and explainable system for automated mental-state assessment and early depression screening.

REFERENCES

[1] Rajawat, Anand Singh, et al. "Fusion fuzzy logic and deep learning for depression detection using facial expressions." *Procedia Computer Science* 218 (2023): 2795-2805.

[2] Xiaoming Cao, Lingling Zhai, Pengpeng Zhai, Fangfei Li, Tao He, Lang He,Deep learning-based depression recognition through facial expression: A systematic review,Neurocomputing,Volume 627,2025,129605,https://doi.org/10.1016/j.neucom.2025.129605.

[3] Qiong Zhao. (2025). Process analysis of facial expressions, movements, and psychological changes in depression based on deep learning algor. *Journal of Biotech Research*. 3285(1944), pp.226-235.

[4] Neha S1, Nivya2, Pooja HC Shekar3, Keerthana S Kumar4, Asha VG5. (2020). Emotion Recognition and Depression Detection using Deep Learning. *International Research Journal of Engineering and Technology*. 7(8), pp.3031 - 3036.

[5] Lee T, Baek S, Lee J, Chung ES, Yun K, Kim TS, Oh J. A Deep Learning Driven Simulation Analysis of the Emotional Profiles of Depression Based on Facial Expression Dynamics. Clin Psychopharmacol Neurosci. 2024 Feb 29;22(1):87-94. doi: 10.9758/cpn.23.1059. Epub 2023 Jun 8. PMID: 38247415; PMCID: PMC10811404.

[6] Lee, Y. S., & Park, W. H. (2022). Diagnosis of depressive disorder model on facial expression based on fast R-CNN. *Diagnostics*, *12*(2), 317.

[7] Almeida, J., & Rodrigues, F. (2021, April). Facial Expression Recognition System for Stress Detection with Deep Learning. In *ICEIS (1)* (pp. 256-263).

[8] Pise, A. A., Alqahtani, M. A., Verma, P., K, P., Karras, D. A., S, P., & Halifa, A. (2022). Methods for facial expression recognition with applications in challenging situations. *Computational intelligence and neuroscience*, *2022*(1), 9261438.

[9] Agung, E.S., Rifai, A.P. & Wijayanto, T. Image-based facial emotion recognition using convolutional neural network on emognition dataset. *Sci Rep* 14, 14429 (2024). https://doi.org/10.1038/s41598-024-65276-x

[10] Liu Y. Using Convolutional Neural Networks for the Assessment Research of Mental Health. Comput Intell Neurosci. 2022 May 9;2022:1636855. doi: 10.1155/2022/1636855. PMID: 35586088; PMCID: PMC9110170.

[11] Hina Tufail, Sehrish Munawar Cheema, Muhammad Ali, Ivan Miguel Pires, Nuno M. Garcia,Depression Detection with Convolutional Neural Networks: A Step Towards Improved Mental Health Care,Procedia Computer Science,Volume 224,2023,Pages 544-549,https://doi.org/10.1016/j.procs.2023.09.079.

[12] Alsadhan, Nasser A. "Image-Based Alzheimer's Disease Detection Using Pretrained Convolutional Neural Network Models." *arXiv preprint arXiv:2502.05815* (2025).

[13] Hafiz, R., Haque, M. R., Rakshit, A., & Uddin, M. S. (2022). Image-based soft drink type classification and dietary assessment system using deep convolutional neural network with transfer learning. *Journal of King Saud University-Computer and Information Sciences*, *34*(5), 1775-1784.

[14] Bhagat, Dhvanil, et al. "Facial emotion recognition (FER) using convolutional neural network (CNN)." Procedia Computer Science 235 (2024): 2079-2089.

[15] Niroshana, S. I., Zhu, X., Nakamura, K., & Chen, W. (2021). A fused-image-based approach to detect obstructive sleep apnea using a single-lead ECG and a 2D convolutional neural network. *Plos one*, *16*(4), e0250618.

[16] Ashawa, M., Owoh, N., Hosseinzadeh, S., & Osamor, J. (2024). Enhanced Image-Based Malware Classification Using Transformer-Based Convolutional Neural Networks (CNNs). *Electronics*, *13*(20), 4081.

[17] Krause, Fernando C., et al. "Facial emotion recognition in major depressive disorder: A meta-analytic review." *Journal of affective disorders* 293 (2021): 320-328.

[18] Fu, Gang, et al. "A method for diagnosing depression: Facial expression mimicry is evaluated by facial expression recognition." *Journal of affective disorders* 323 (2023): 809-818.

[19] Gao, Zhiyun, Wentao Zhao, Sha Liu, Zhifen Liu, Chengxiang Yang, and Yong Xu. "Facial emotion recognition in schizophrenia." *Frontiers in psychiatry* 12 (2021): 633717.

[20] M. Li, J. Zhang, J. Song, Z. Li and S. Lu, "A Clinical-Oriented Non-Severe Depression Diagnosis Method Based on Cognitive Behavior of Emotional Conflict," in *IEEE Transactions on Computational Social Systems*, vol. 10, no. 1, pp. 131-141, Feb. 2023, doi: 10.1109/TCSS.2022.3152091.

[21] Khan, A. R. (2022). Facial emotion recognition using conventional machine learning and deep learning methods: current achievements, analysis and remaining challenges. *Information*, *13*(6), 268.

[22] Chowanda, A. Separable convolutional neural networks for facial expressions recognition. *J Big Data* 8, 132 (2021). https://doi.org/10.1186/s40537-021-00522-x

[23] Aleem, S., Huda, N. U., Amin, R., Khalid, S., Alshamrani, S. S., & Alshehri, A. (2022). Machine learning algorithms for depression: diagnosis, insights, and research directions. *Electronics*, *11*(7), 1111.

[24] Squires, M., Tao, X., Elangovan, S. *et al.* Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment. *Brain Inf.* 10, 10 (2023). https://doi.org/10.1186/s40708-023-00188-6

[25] F. Moreno, S. Alghowinem, H. W. Park and C. Breazeal, "Expresso-AI: An Explainable Video-Based Deep Learning Models for Depression Diagnosis," *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, Cambridge, MA, USA, 2023, pp. 1-8, doi: 10.1109/ACII59096.2023.10388143.

[26] Kerz, Elma, et al. "Toward explainable AI (XAI) for mental health detection based on language behavior." *Frontiers in psychiatry* 14 (2023): 1219479.

[27] Al Masud, Gazi Hasan, et al. "Effective depression detection and interpretation: Integrating machine learning, deep learning, language models, and explainable AI." *Array* (2025): 100375.

[28] El-Sappagh, S., Alonso, J.M., Islam, S.M.R. *et al.* A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Sci Rep* 11, 2660 (2021). https://doi.org/10.1038/s41598-021-82098-3

[29] M. Ahmad Wani, M. A. ELAffendi, K. A. Shakil, A. Shariq Imran and A. A. Abd El-Latif, "Depression Screening in Humans With AI and Deep Learning Techniques," in *IEEE Transactions on Computational Social Systems*, vol. 10, no. 4, pp. 2074-2089, Aug. 2023, doi: 10.1109/TCSS.2022.3200213.

[30] M. A. Hossain, A. K. M. M. Islam, S. Islam, S. Shatabda and A. Ahmed, "Symptom Based Explainable Artificial Intelligence Model for Leukemia Detection," in *IEEE Access*, vol. 10, pp. 57283-57298, 2022, doi: 10.1109/ACCESS.2022.3176274

[31] Joyce, D.W., Kormilitzin, A., Smith, K.A. *et al.* Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digit. Med.* 6, 6 (2023). https://doi.org/10.1038/s41746-023-00751-9

[32] Sun, Q., Akman, A., & Schuller, B. W. (2025). Explainable artificial intelligence for medical applications: A review. *ACM Transactions on Computing for Healthcare*, *6*(2), 1-31.

[33] Zhang, Z., Zhang, S., Ni, D., Wei, Z., Yang, K., Jin, S., ... & Wang, J. (2024). Multimodal sensing for depression risk detection: Integrating audio, video, and text data. *Sensors*, *24*(12), 3714.

[34] Zhang, L., Zhang, S., Zhang, X., & Zhao, Y. (2025). A Multimodal Artificial Intelligence Model for Depression Severity Detection Based on Audio and Video Signals. *Electronics*, *14*(7), 1464.

[35] Yoon, J., Kang, C., Kim, S., & Han, J. (2022, June). D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 11, pp. 12226-12234).

[36] Wang, C., Liang, L., Liu, X., Lu, Y., Shen, J., Luo, H., & Xie, W. (2021, November). Multimodal fusion diagnosis of depression and anxiety based on face video. In *2021 IEEE International Conference on Medical Imaging Physics and Engineering (ICMIPE)* (pp. 1-7). IEEE.

[37] Gimeno-Gómez, D., Bucur, AM., Cosma, A., Martínez-Hinarejos, CD., Rosso, P. (2024). Reading Between the Frames: Multi-modal Depression Detection in Videos from Non-verbal Cues. In: Goharian, N., *et al.* Advances in Information Retrieval. ECIR 2024. Lecture Notes in Computer Science, vol 14608. Springer, Cham. https://doi.org/10.1007/978-3-031-56027-9_12

[38] Gilanie, G., Asghar, M., Qamar, A. M., Ullah, H., Khan, R. U., Aslam, N., & Khan, I. U. (2022). An Automated and Real-time Approach of Depression Detection from Facial Micro-expressions. *Computers, Materials & Continua*, *73*(2).

[39] G. Tiwary, S. Chauhan and K. K. Goyal, "Multimodal Depression Detection Using Audio Visual Cues," *2023 International Conference on Computer Science and Emerging Technologies (CSET)*, Bangalore, India, 2023, pp. 1-5, doi: 10.1109/CSET58993.2023.10346770.

[40] Park, J., & Moon, N. (2022). Design and implementation of attention depression detection model based on multimodal analysis. *Sustainability*, *14*(6), 3569.

[41] Gratch, J., Artstein, R., Lucas, G.M., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D.R. (2014) *The Distress Analysis Interview Corpus of Human and Computer Interviews*. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), pp. 3123–3128.